Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Proper Proxy Scoring Rules

Jens Witkowski
ETH Zurich
jensw@inf.ethz.ch

Pavel Atanasov
Pytho LLC
pavel@pytho.io

Lyle H. Ungar
University of Pennsylvania
ungar@cis.upenn.edu

Andreas Krause
ETH Zurich
krausea@ethz.ch

Abstract

Proper scoring rules can be used to incentivize a forecaster to truthfully report her private beliefs about the probabilities of future events and to evaluate the relative accuracy of forecasters. While standard scoring rules can score forecasts only once the associated events have been resolved, many applications would benefit from instant access to proper scores. In forecast aggregation, for example, it is known that using weighted averages, where more weight is put on more accurate forecasters, outperforms simple averaging of forecasts.

We introduce *proxy scoring rules*, which generalize proper scoring rules and, given access to an appropriate proxy, allow for immediate scoring of probabilistic forecasts. In particular, we suggest a proxy-scoring generalization of the popular quadratic scoring rule, and characterize its incentive and accuracy evaluation properties theoretically. Moreover, we thoroughly evaluate it experimentally using data from a large real world geopolitical forecasting tournament, and show that it is competitive with proper scoring rules when the number of questions is small.

1 Introduction

We study the problem of evaluating the accuracy of probabilistic forecasts. This is typically done using proper scoring rules (Brier 1950; Good 1952; Gneiting and Raftery 2007), which are used in two ways. First, they are used as incentive mechanisms that truthfully elicit an agent's private belief about the outcome of a future event. Second, they are used as an accuracy evaluation mechanism to judge the accuracy of probabilistic forecasts even if incentives are not a concern. The incentive and the evaluation aspect of proper scoring rules are two sides of the same coin, and often both are required. For example, proper scoring rules have a long history in meteorology for comparing the relative accuracy of competing weather forecasters (Brier 1950; Good 1952). While the primary goal is evaluating accuracy, it is important that weather services are not incentivized to misreport their true estimates. Another example is geopolitical forecasting, where forecasters are asked to report probabilistic estimates of the likelihood of geopolitical events, such as the probability of a military coup in Yemen (Atanasov et al. 2016). Here, proper scoring is important to incentivize forecasters to invest effort and report truthfully, and to judge the relative accuracy of forecasts.

For proper scoring rules to be applicable, the forecasting system requires the resolved outcomes of the events that forecasters reported on. In this paper, we introduce proxy scoring rules, which generalize proper scoring rules and, given access to an appropriate proxy, allow for immediate scoring of probabilistic forecasts, i.e., right after elicitation and before outcome resolution. This is important for a number of applications. In forecast aggregation, for example, it is known that using weighted averages, where more weight is put on more accurate forecasters, outperforms simple averaging of forecasts (Atanasov et al. 2016). To make this work with proper scoring rules, one requires early-closing questions, which can then inform the weighting of forecasters for other questions. Often, however, many questions close at the same time. Consider for example forecasting the majority party in each of the 50 US states in a presidential election. Voting closes on the same day in all states, and so the system does not obtain forecaster accuracy information in time to improve aggregation through forecaster weighting. Or consider the challenge of scoring long-term questions, such as "Will humanity have traveled to Mars by 2030?" Applying proper scoring rules to such questions would require forecasters to wait for payment until 2030. It is unlikely that forecasters have the patience and trust in the continued existence of the system that this would require. Or consider the problem of hiring intelligence agents based on their accuracy in forecasting geopolitical events. It is not feasible to wait until a sufficient number of questions have been resolved before making a hiring decision. Instead, the intelligence agency requires immediate feedback on the accuracy of the candidate.

In addition to proper scoring rules, proxy scoring rules are related to peer prediction mechanisms (Miller, Resnick, and Zeckhauser 2005; Prelec 2004; Jurca 2007; Witkowski 2014; Waggoner and Chen 2014), where the system seeks to incentivize agents to report private information without being able to verify the truthfulness of their reports. Peer prediction is related to our approach in that agents are not scored against an outcome of ground truth but a proxy event (in the case of peer prediction, this is another agent's report). The boundary between this work and peer prediction

is smooth and partly depends on the proxy that is being used. For example, the extremized mean proxy that we suggest in Section 4 bears a lot of resemblance to peer prediction since it is an aggregate of other agents' reports. One difference to peer prediction is that, at the heart of proper and proxy scoring is the elicitation and scoring of probabilistic forecasts, whereas peer prediction is primarily concerned with the elicitation of informative signals. While Bayesian Truth Serum mechanisms (Prelec 2004; Witkowski and Parkes 2012; Radanovic and Faltings 2013) also elicit a probabilistic report, the probabilistic report is only a means towards eliciting the signal. The proxy scoring framework can be viewed as providing a link between the literature on proper scoring rules and peer prediction.

Chakraborty and Das (2016) study a two-trader prediction market model, where the outcome of the event is decided by a vote of the traders themselves. Similarly, Freeman, Lahaie, and Pennock (2017) study a model where a prediction market is followed by a peer prediction mechanism to determine the outcome. While that literature also combines forecasting with peer prediction, proxy scoring rules generalize proper scoring rules and, as such, proxy scoring rules only ask agents for one report (a probabilistic forecast) instead of having two reporting stages (a forecasting and an outcome determination stage). Moreover, just as proper scoring rules, proxy scoring rules can be used both to provide proper incentives and to estimate forecaster accuracy.

The remainder of the paper is organized as follows. First, we review proper scoring rules in Section 2. In Section 3, we then introduce a new class of scoring rule called *proxy scoring rules*, which, given access to an appropriate proxy, allow for the immediate scoring of probabilistic forecasts. We characterize the conditions on the rule and the proxy that need to hold for properness and provide concrete examples. Moreover, we show that the class of proper proxy scoring rules contains proper scoring rules as a special case. In Section 4, we show how the theoretical concepts of Section 3 translate to a concrete application. After introducing the data set, we introduce an example proxy that we then thoroughly evaluate on a large, real-world data set from a geopolitical forecasting tournament. In particular, we show that a generalization of the quadratic proper scoring rule together with the example proxy predicts out of sample forecaster accuracy almost as good as the quadratic proper scoring rule that has access to true event outcomes.

2 Proper Scoring Rules

Proper scoring rules are used for two purposes: as an incentive mechanism, which incentivizes rational forecasters to truthfully report their private, probabilistic beliefs about the likelihood of a future event, and as an evaluation mechanism, which estimates the relative accuracy of forecasts.

Consider first a single forecaster and let $p \in [0,1]$ be her probabilistic belief that $\omega = 1$. The scoring proceeds as follows: first, the system (center) asks the forecaster for her belief report $y \in [0,1]$. Second, an event $\omega \in \{0,1\}$ materializes (observed by the center) and, third, the center pays the forecaster the payment $R(y,\omega)$.

Definition 1 (Scoring Rule)**.** *Given outcome $\omega \in \{0,1\}$ and report $y \in [0,1]$ in regard to the probability that $\omega = 1$, a scoring rule $R(y,\omega) \in \mathbb{R} \cup \{-\infty, +\infty\}$ assigns a score based on report y and the outcome ω that occurs.*

Definition 2 (Strictly Proper Scoring Rule)**.** *A scoring rule is* proper *if a forecaster maximizes her expected score by truthfully reporting her belief $p \in [0,1]$, and is* strictly proper *if the truthful report is the only report that maximizes the forecaster's expected score.*

Definition 2 is phrased in an incentive spirit. An equivalent definition in the evaluation spirit is the following:

Definition 3 (Strictly Proper Scoring Rule)**.** *Let $\theta = \Pr(\omega = 1)$ be the true probability of the event occurring and let there be two forecasts $y = \theta$ and $y' \neq \theta$. Then scoring rule R is* proper *if $\mathbf{E}_\omega[R(y,\omega)] \geq \mathbf{E}_\omega[R(y',\omega)]$ for all $\theta \in [0,1]$, and* strictly proper *if the inequality is strict.*

There exist infinitely many proper scoring rules since any (strictly) convex function corresponds to a (strictly) proper scoring rule (Gneiting and Raftery 2007, Theorem 1). Where these rules differ from one another is in how imperfect forecasts (or: forecasters) are scored. For example, let $\theta = 0.7$ be the true probability and consider forecasts $y = 0.79$ and $y' = 0.6$. Since neither y nor y' is the true probability, strict properness does not dictate which forecast shall receive a higher expected score. This is specified only through the particular choice of proper scoring rule. In this paper, we focus on the *quadratic scoring rule* (Brier 1950).

Definition 4. *The* quadratic scoring rule *normalized to yield scores in the interval $[0,1]$ is $R_q(y,\omega) = 1 - (y-\omega)^2$.*

Proposition 1. *(Brier 1950) The quadratic scoring rule R_q is strictly proper.*

(We prove a generalization of this result in Section 3.)

For the quadratic rule, the expected score difference between the highest possible score (for reporting the true probability θ) and reporting $y \in [0,1]$ is $(y-\theta)^2$ (Selten 1998, p.47f). Applying it to the above example, the expected score for report $y = 0.79$ would thus be higher than the expected score for report $y' = 0.6$ since $(0.79-0.7)^2 < (0.6-0.7)^2$.[1] In fact, the quadratic rule is the only strictly proper scoring rule, where the expected loss is a function of only $y - \theta$, i.e., the difference between the forecast and the true probability (Savage 1971, p.787f). Moreover, the quadratic scoring rule is convenient for payments since it is bounded (here scaled to be in $[0,1]$) and non-negative.

In practice, forecasters often report on more than one event, in which case, scores are typically averaged over questions (e. g., Brier 1950, Gneiting and Raftery 2007, Atanasov et al. 2016). With the quadratic scoring rule, for example, a forecaster reporting $y_1, \ldots, y_n \in [0,1]^n$ on n different questions is assigned score $\frac{1}{n}\sum_{i=1}^{n} R_q(y_i, \omega_i)$, where ω_i denotes the outcome of the ith event. Without any assumptions on how the events are related, comparing average scores is only meaningful if forecasters report on the

[1] Note that this decision is not self-evident. The logarithmic proper scoring rule (Good 1952), for example, gives higher expected score to $y' = 0.6$ than $y = 0.79$ since $0.7\ln(0.6) + 0.3\ln(0.4) > 0.7\ln(0.79) + 0.3\ln(0.21)$.

same set of questions. This is because the highest possible expected score (of a perfect forecast) depends on θ. For example, for the quadratic rule, if $\theta = 0.5$, a perfect forecast of $y = \theta = 0.5$ yields an expected score of 0.75, whereas for $\theta = 0.1$, a perfect forecast of $y = \theta = 0.1$, yields an expected score of 0.91. For a more in-depth treatment of proper scoring rules, we refer to the article by Gneiting and Raftery (2007).

3 Proxy Scoring Rules

In this section, we introduce *proxy scoring rules*, a generalization of proper scoring rules, which allow for the immediate scoring of probabilistic forecasts.

3.1 Model

As in Section 2, let again θ denote the probability that $\omega = 1$ and let y be a forecaster's report. In contrast to the previous section, instead of eventually observing the event's outcome ω, the center now has access to a *proxy* $\hat{\Theta}$, which is a random variable conditioned on θ, i.e., $\hat{\Theta} \sim \Pr(\hat{\Theta} \,|\Theta = \theta)$. At this point, we are agnostic as to where such a proxy may come from. As we will see in Section 4.2, one possibility is to estimate a proxy from the forecasters' forecasts themselves, e.g. using a transformation of the average forecast on a question. We assume that the forecaster knows θ and the distribution of random variable $\hat{\Theta}$, but nothing about its realization $\hat{\theta}$. Note that these knowledge assumptions are with respect to the forecaster's subjective probability. Of course, a forecaster may actually not know the true θ but instead have some subjective belief p about the event occurring. The objective of proper scoring can be cast as: "Assume you knew θ, then you should maximize your expected score reporting θ."

3.2 Proper Proxy Scoring Rules

Definition 5 (Proxy Scoring Rule)**.** *Let $y \in [0, 1]$ be a belief report in regard to the probability that $\omega = 1$, and let $\hat{\theta} \in [0, 1]$ be the proxy forecast. A proxy scoring rule $R(y, \hat{\theta}) \in \mathbb{R} \cup \{-\infty, +\infty\}$ assigns a score based on belief report y and proxy forecast $\hat{\theta}$.*

The crucial difference to proper scoring rules from Definition 1 is that a proxy scoring rule cannot use the actual outcome ω stemming from θ but only has access to a proxy forecast $\hat{\theta}$ stemming from proxy $\hat{\Theta}$.

Definition 6 (Strictly Proper Proxy Scoring Rule)**.** *Let $\theta = \Pr(\omega = 1)$ be the true probability of the event occuring and let there be two forecasts $y = \theta$ and $y' \neq \theta$. Then proxy scoring rule $R(y, \hat{\theta})$ is* proper *if $\mathbf{E}_{\hat{\Theta}}[R(y, \hat{\Theta}) \mid \Theta = \theta] \geq \mathbf{E}_{\hat{\Theta}}[R(y', \hat{\Theta}) \mid \Theta = \theta]$ for all $\theta \in [0, 1]$, and* strictly proper *if the inequality is strict.*

Before we characterize conditions on $\hat{\Theta}$ and $R(y, \hat{\theta})$ under which $R(y, \hat{\theta})$ is strictly proper, consider first a special case: intuitively, the goal of both proper scoring rules and proper proxy scoring rules is to elicit θ, i.e., the expectation of ω. In proxy scoring, we do not have access to ω directly but get around this restriction through access to a proxy, which is statistically linked to the true distribution θ. For example, assume we knew the expectation of $\hat{\Theta}$ and that it equals the expectation of the true distribution, i.e., $\mathbf{E}[\hat{\Theta} \,|\theta] = \theta$. If we had a way to incentivize the forecaster to report the mean of $\Pr(\hat{\Theta} \,|\theta)$, then we would be able to incentivize reporting θ (or the forecaster's best estimate of θ) implicitly.

To characterize the more general case, we now turn to the literature on property elicitation, where one wishes to extract a particular function, or *property*, of an agent's belief using a scoring rule with access to a single sample from the true distribution (which, in our case, will be the proxy) (Lambert, Pennock, and Shoham 2008; Frongillo and Kash 2014). We present the following definitions for the special case that we need in this paper.

Definition 7 (Property)**.** *(Lambert, Pennock, and Shoham 2008) We call $\Gamma : [0, 1] \to [0, 1]$ a* property *of distribution $\mathcal{P}$.*

Examples of properties are the mean, the median, or the variance.

Definition 8 (Property Scoring Rule)**.** *(Lambert, Pennock, and Shoham 2008) Let $y \in [0, 1]$ be a report in regard to $\Gamma(\mathcal{P})$ and let x be a sample from $\mathcal{P}$. Property scoring rule $R(y, x) \in \mathbb{R} \cup \{-\infty, +\infty\}$ assigns a score based on belief report y and sample x.*

Definition 9 (Strictly Proper Property Scoring Rule)**.** *(Lambert, Pennock, and Shoham 2008) Let $y = \Gamma(\mathcal{P})$ and $y' \neq \Gamma(\mathcal{P})$. Then property scoring rule R is* proper *if $\mathbf{E}_{\mathcal{P}}[R(y, x)] \geq \mathbf{E}_{\mathcal{P}}[R(y', x)]$ for all $\mathcal{P} \in [0, 1]$, and* strictly proper *if the inequality is strict.*

An example of a property scoring rule that is proper for the mean is the quadratic rule R_q (also see Lemma 5).

Theorem 2. *Proxy scoring rule $R(y, \hat{\theta})$ is (strictly) proper if property scoring rule R is (strictly) proper for Γ and $\Gamma(\Pr(\hat{\Theta} \,|\theta)) = \theta$ for all θ.*

Proof. The statement follows by construction: property scoring rule R is (strictly) proper for Γ and $\Gamma(\Pr(\hat{\Theta} \,|\theta)) = \theta$. This means that for $y = \theta$ and $y' \neq \theta$, it holds that $\mathbf{E}_{\hat{\Theta}}[R(y, \hat{\Theta}) \mid \Theta = \theta] \geq \mathbf{E}_{\hat{\Theta}}[R(y', \hat{\Theta}) \mid \Theta = \theta]$ for all $\theta \in [0, 1]$, where the inequality is strict if R is strictly proper for Γ. □

Corollary 3. *Proxy scoring rule $R(y, \hat{\theta})$ is strictly proper if $\mathbf{E}[\hat{\Theta} \mid \theta] = \theta$, and property scoring rule R is strictly proper for the mean, i.e., $\Gamma(\Pr(\hat{\Theta} \,|\theta)) = \mathbf{E}[\hat{\Theta} \,|\theta]$.*

Corollary 3 is Theorem 2 for Γ being the mean. To see how Theorem 2 is more general than Corollary 3, consider the case where Γ is the median instead of the mean. This also results in a strictly proper proxy scoring rule if the median of $\Pr(\hat{\Theta} \,|\theta)$ is equal to θ and property scoring rule R is strictly proper for the median.

3.3 Proper Scoring Rules as Special Case

Theorem 4 establishes that the class of proper proxy scoring rules contains proper scoring rules as a special case.

Theorem 4. *Let $R(y,\omega)$ be a (strictly) proper scoring rule. Then proxy scoring rule $R(y,\hat{\theta})$ with*

$$\hat{\Theta} \sim \Pr(\hat{\Theta}\,|\theta) = \begin{cases} 0 & \textit{with } 1-\theta \\ 1 & \textit{with } \theta \end{cases}$$

is (strictly) proper and yields the same payment distribution as $R(y,\omega)$ given y and θ. That is, random variables $R(y,\hat{\theta})$ and $R(y,\omega)$ are identical for all $\theta, y \in [0,1]$.

Proof. The statement follows directly from the fact that random variables ω and $\hat{\Theta}$ are identically distributed according to $\Pr(\hat{\Theta}\,|\theta) = \Pr(\omega)$. □

Note that without further knowledge about the dependency of $\hat{\Theta}$ and ω, one cannot make the statement that $R(y,\hat{\theta}) = R(y,\omega)$ since drawing one sample each from two identically distributed random variables does not mean that these samples are identical (even if the probabilities are). However, Theorem 4 is stronger than stating that only the expected payments of both rules agree since that would allow for the payment distributions to be different as long as their expectations are identical.

3.4 The Quadratic Proxy Scoring Rule

Definition 10 (Quadratic Proxy Scoring Rule)**.** *The* Quadratic Proxy Scoring Rule *is defined as*

$$R_q(y,\hat{\theta}) = 1 - (y-\hat{\theta})^2.$$

It is easy to see that the quadratic proxy scoring rule is a generalization of the quadratic proper scoring rule.

Lemma 5. *(Brier 1950; Savage 1971) $R_q(y,\cdot)$ is strictly proper for the mean.*

Theorem 6. *The quadratic proxy scoring rule with proxy $\hat{\Theta}$ is strictly proper iff $\mathbf{E}[\hat{\Theta} \mid \theta] = \theta$.*

Proof. The statement follows directly from Theorem 2 and Lemma 5. □

We refer to proxies satisfying $\mathbf{E}[\hat{\Theta} \mid \theta] = \theta$ as *unbiased*. Observe that unbiasedness only requires the mean of the proxy to be the true probability; in particular, it is not required for properness that any sample from the proxy ever coincides with the true probability.

4 Experimental Evaluation

In this section, we evaluate the quadratic proxy scoring rule with an example proxy experimentally using a real-world forecasting data set.

4.1 Good Judgment Data Set

The data set we use is from the *Good Judgment Project*, a research and development project that provides probabilistic forecasts of geopolitical events to the United States intelligence community. We use data from the third year, which includes questions started on or after August 1, 2013 and were scheduled to close on or before June 1, 2014. This original data set comprises 515 forecasters (with and without prediction training), who made more than 48,000 forecasts on 100 questions. An example question is "Will Angela Merkel win the next election for Chancellor of Germany?" Forecasters self selected the questions they want to report on. For more details, including how the incentive issues that are inherent in self selection were addressed, we refer to the work by Atanasov et al. (2016).

As expected from a real-world application, this dynamic environment has a fair amount of complexity in a number of dimensions. In particular, forecasters are free to update their forecasts over time, only very few forecasters report on all questions, and while most questions are binary, some have more than two possible outcomes. To obtain a (forecast, outcome) tuple with binary outcome for all questions, we first map every non-binary question to a binary question by interpreting the first of the possible outcomes as "event occurred" and all others as "event did not occur." We then, for each question, subset to those forecasters who made at least one forecast for that question in the first 7 days after the question started, and define a forecaster's forecast for that question as her average forecast within those first 7 days. (7 days were chosen because there is a rush of forecasting activity in the first week of a question.) We dropped 2 questions in the original data set, which lasted for 1 day and 7 days, respectively, after the question was posed. After this transformation of the original data set, we are left with 426 forecasters, who made 5728 forecasts on 98 questions.

4.2 The Extremized Mean Proxy

From Theorem 6 we know that the quadratic proxy scoring rule is proper with an unbiased proxy. What has not been discussed so far is where such a proxy may come from. One possibility is to use an aggregate of the forecasts themselves, such as the mean or the median of all forecasts on a given question. Two challenges with this kind of proxy shall be noted here: first, any such proxy will be slightly biased in practice and thus the proxy scoring rule will not be perfectly proper. Note that to take advantage of this, strategic forecasters would need to know the proxy's bias, which is perhaps unlikely. Second, when using such a proxy for incentives instead of evaluating forecaster accuracy, the resulting scheme shares many similarities with peer prediction mechanisms (Miller, Resnick, and Zeckhauser 2005; Prelec 2004; Jurca 2007; Witkowski 2014), including the existence of non-truthful equilibria (Waggoner and Chen 2014; Shnayder, Frongillo, and Parkes 2016). The empirical evidence as to whether or not the existence of non-truthful equilibria are a problem in practice are mixed (John, Loewenstein, and Prelec 2012; Gao, Mao, and Chen 2014; Rigol 2016). While we focus on the accuracy evaluation property of proxy scoring in this section, it is an interesting direction for future work to experimentally evaluate the incentives of proxy scoring when the computation of the proxy itself is using the forecasters' forecasts.

One of the best performing simple aggregators (i.e., not weighting forecasters by their frequency or magnitude of forecasts) of the Good Judgment Project (in terms of quadratic scoring rule) was the extremized mean.

Definition 11. *(Atanasov et al. 2016) Let $\bar{y}_i$ be the mean of forecasts for question i. The* extremized mean *is given by*

$$\hat{\theta}_{i,\alpha} = \frac{{\bar{y}_i}^\alpha}{{\bar{y}_i}^\alpha + (1-\bar{y}_i)^\alpha},$$

where $\alpha \geq 1$ is an extremizing parameter of the aggregator.

It is easy to see that for $\alpha = 1$, this is just the simple mean, whereas for $\alpha > 1$, any $\bar{y}_i > 0.5$ will be pushed towards 1 and any $\bar{y}_i < 0.5$ will be pushed towards 0. The optimal α for the original dynamic Good Judgment data set, which was optimized out of sample from earlier seasons, is $\alpha = 2$ (Atanasov et al. 2016), and so we are also using $\alpha = 2$ in our experiments. For intuition as to why the extremized mean is a good aggregator, see Appendix A.

It is worth emphasizing that the aggregation of forecasts is not the focus of this paper. Instead, we use the extremized mean of forecasts (an aggregation algorithm; Definition 11) as the proxy in the quadratic proxy scoring rule (Definition 10). Just as with standard proper scoring rules, our two objectives are to incentivize truthtelling and to estimate the relative accuracy of forecasters. (Both properties are implied by a rule being *proper*.) It shall also be noted here that when computing the proxy, we use the same proxy for all forecasters and thus ignore a forecaster's influence on the computation of the proxy through her own forecast. Of course, one can compute the proxy (e. g. the extremized mean) leaving out the forecasters whose accuracies are to be evaluated and run this computation for every forecaster pair but given the large number of forecasters in our data set, this would not make a qualitative difference to the results.

4.3 Estimation Procedure

Both proper scoring and proxy scoring address the problem of deciding which of two forecasters is better, where better means higher expected quadratic scoring rule score on the same set of questions. The goal of this section is to compare the performance of the quadratic scoring rule with access to true event outcomes with the performance of the quadratic proxy scoring rule, which only has access to the extremized mean proxy. Answering this problem using the data described in Section 4.1 is difficult because neither are the forecasters' questions sampled i.i.d. nor do we have infinite data. To guarantee that different forecasters' scores are comparable, we only compare average scores on questions that both forecasters reported on.

We subset to forecaster pairs with at least 60 questions in common (of which there are 210) and, for each forecaster pair, randomly sample two sets of questions, which we refer to as the selection set and the validation set. The validation set has size 30 and is scored using the quadratic proper scoring rule from Definition 4 with access to the corresponding 30 event outcomes. Whichever of the two forecasters obtains a higher score on the validation set is considered its winner. (We will come back to the problem that this is not ground truth.) The selection set's size goes from 1 to 30 and is scored using both the proper quadratic score (with access to the selection set's outcomes) and the quadratic proxy score, where for each question, the proxy forecast is the extremized mean of that question. Both methods call one of the two forecasters the selection set winner. A method receives a point if and only if its selection set winner agrees with the validation set winner. For each of the 210 forecaster pairs, we sample the validation set 10 times and, within each of those, also sample the selection set 10 times for every size from 1 to 30. We estimate the methods' probabilities of agreeing with the validation set winner as the fraction of points over the number of samples, which provides us with a relative performance measure of the two methods.

Agreement vs Being Correct The validation set winner is itself only a noisy predictor of who is the better forecaster. Ideally, we not only have such a relative comparison but obtain estimates of the true probability that a given method is predicting correctly who is the better forecaster. To obtain this estimate, we first estimate that the validation set winner is correct. We proceed as follows: subsetting again to forecaster pairs with at least 60 questions in common, we sample two equally-sized sets of size 30. This time, we score both sets using only the proper quadratic score with access to the outcomes in the respective sets. Both sets call a winner and obtain a point if they agree. We then again divide the number of points by the number of samples to obtain a good estimate of the probability that the winner of the two sets agree.

Observe that the probability of correctly predicting the better forecaster is the same for both sets since the two sets are statistically identical. Let c_v be this probability that any one of the sets is predicting the better forecaster correctly, and let a_v be the probability of the two sets' winners agreeing. a_v can then be expressed as $a_v = c_v^2 + (1-c_v)^2$. After simple algebra and assuming that $a_v > 0.5$, we obtain $c_v = 0.5 + \frac{\sqrt{2a_v - 1}}{2}$, which gives us a way to estimate that the validation set winner correctly predicts the better forecaster. For 1000 samples, we obtain estimates $a_v = 0.71$ and $c_v = 0.82$.

With c_v in hand, we can use a similar procedure to estimate the probability that the selection set winner correctly predicts the better forecaster. Let c_s be this probability, and let a_s be the probability that the selection set winner agrees with the validation set winner. a_s can then be expressed as $a_s = c_s \cdot c_v + (1-c_s)(1-c_v)$. Since we know a_s and have just estimated c_v, we can solve for c_s and, assuming $c_v > 0.5$, we obtain $c_s = \frac{c_v + a_s - 1}{2c_v - 1}$.

4.4 Results

The experimental results are shown in Figure 1. Slightly abusing definitions, we will refer to the two methods as proper scoring rule and proxy scoring rule, respectively, in an effort to avoid confusion. The "proper scoring rule" is the quadratic scoring rule (Definition 4) using resolved outcomes of the respective set. The "proxy scoring rule" is the quadratic proxy scoring rule (Definition 10) using the extremized mean proxy (Definition 11) on the respective set. (Of course, every proper scoring rule can be written as a proper proxy scoring rule (Theorem 4), so that proxy does not imply improper and proper does not exclude proxy.) Note that proper scoring may not be realistic in practice since it requires that questions have already been resolved.

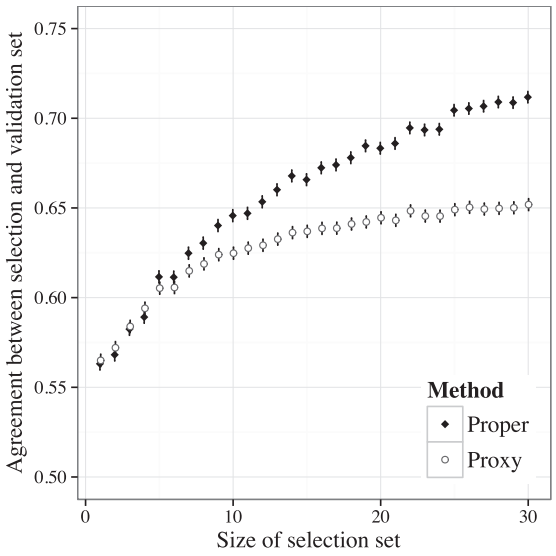

Figure 1: Plotting the probability of agreement between the selection and the validation set with standard errors. The validation set contains 30 questions and the selection set sizes range from 1 to 30. Proper scoring requires resolved questions in the selection set, whereas proxy scoring only uses an aggregate of the forecasts on the selection set.

While not perfectly accurate in predicting the better forecaster, the validation set winner is proper-scored using the outcomes of 30 questions and is thus a decent approximation of ground truth. As mentioned in Section 4.3, we estimate the probability of it being correct to be $c_v = 0.82$. Therefore, the larger the agreement with the validation set, the more accurate is the method at predicting who is the better forecaster. Unsurprisingly, both methods increase agreement as the size of the selection set increases: the more scored questions there are in the selection set, the more information it contains, and so both methods do better. Keeping the size of the validation set fixed, $c_v = 0.82$ is an upper bound on the agreement probability even for proper scores and infinitely large selection sets. The reason is that while the probability that proper scoring will correctly call the better forecaster goes to 1 as the number of questions in the selection set increases, the validation set winner will still be correct only with probability $c_v = 0.82$, and thus the probability of agreement cannot be higher.

Proxy scoring, however, seems to converge to a level below 0.82. For a selection set size of 30, we obtain an agreement probability of $a_s = 0.65$, which, using the procedure introduced in Section 4.3, we estimate to correspond to a probability of calling the correct winner of $c_s = 0.73$. Note that this is still around 89% of the probability that proper scoring achieves ($0.89 \simeq 0.73/0.82$). Starting with 5 questions in the selection set, proxy scoring is increasing slower than proper scoring. This slope to a lower level than proper scoring suggests that the extremized mean proxy is not perfectly unbiased. For the first 4 questions, however, proxy scoring slightly outperforms proper scoring. Presumably, this is the case because average proxy scores have less variance than average proper scores with such few questions. Consider the most extreme case of only one question in the selection set: the proper score calls that forecaster the winner who is closer to 0 or 1, depending on the actual outcome. The proxy score is less extreme and calls that forecaster the winner who is closer to the proxy forecast, which will typically be a number strictly in between 0 and 1.

An interesting additional observation is that for validation and selection set sizes of 30 each, the in sample prediction of proxy scoring (i.e., computing the average proxy score on the 30 validation set questions) is as good in predicting average proper score on the validation set as the out of sample prediction of proper scoring (both agree with the validation set winner with probability 0.71). This is interesting because it depends on the use of proxy scoring as to whether in or out of sample is the right evaluation method. Out of sample is the correct method if one wants to estimate which forecaster will do better on other questions in the future. In sample is appropriate when using proxy scoring in aggregation algorithms, where one wants to put more weight on forecasters who are doing well on the currently open questions.

5 Conclusion

We introduced a new class of scoring rules generalizing proper scoring rules to settings, where the center does not have access to samples from the true distribution. We characterized conditions that allow for proper proxy scoring, and experimentally evaluated the performance of the quadratic proxy scoring rule with the extremized mean proxy on a real-world geopolitical forecasting data set.

We believe an exciting direction for future work is to use proxy scores in aggregation algorithms. It has been shown that forecast aggregation is improved relative to unweighted aggregation when algorithms put more weight on those forecasters who have higher average quadratic score on already resolved questions (Atanasov et al. 2016). Of course, using proxy scores instead of scores using already closed questions is promising because forecasting systems do not always have access to resolved questions. Moreover, and perhaps most interestingly, for aggregation, one can use the proxy scores *in sample*, i.e., on the questions that the aggregation algorithm is eventually scored against. Here, our experiments give an agreement probability of 0.71, which is the same as the agreement probability of out of sample proper scoring with the quadratic scoring rule. That is, using the average proxy score on the 30 questions that have just been posed is as predictive of who is eventually obtaining a higher score on these questions as using the average proper score on 30 out-of-sample resolved questions.

Another interesting direction for future work is equilibrium selection when using an aggregate of other forecasters' forecasts as the proxy. The peer prediction community has recently made progress in making coordination on non-truthful equilibria "less likely" (e.g. Dasgupta and Ghosh 2013; Shnayder, Frongillo, and Parkes 2016), and it will be interesting to see how these techniques can be adapted to proxy scoring.

Acknowledgments

We thank the anonymous reviewers, Michael M. Bishop, Angela Minster, Rafael Frongillo, and all members of the Good Judgment Project for helpful discussions. This work is supported by the Nano-Tera.ch program as part of the Opensense II project.

References

Atanasov, P.; Rescober, P.; Stone, E.; Servan-Schreiber, E.; Tetlock, P. E.; Ungar, L.; and Mellers, B. 2016. Distilling the Wisdom of Crowds: Prediction Markets versus Prediction Polls. *Management Science*. forthcoming.

Baron, J.; Mellers, B. A.; Tetlock, P. E.; Stone, E.; and Ungar, L. H. 2014. Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis* 11(2):133–145.

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78(1):1–3.

Chakraborty, M., and Das, S. 2016. Trading On A Rigged Game: Outcome Manipulation In Prediction Markets. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, 158–164.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22nd ACM International World Wide Web Conference (WWW'13)*, 319–330.

Freeman, R.; Lahaie, S.; and Pennock, D. 2017. Crowdsourced Outcome Determination in Prediction Markets. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. forthcoming.

Frongillo, R., and Kash, I. 2014. General truthfulness characterizations via convex analysis. In *Web and Internet Economics*, 354–370. Springer.

Gao, A.; Mao, A.; and Chen, Y. 2014. Trick or Treat: Putting Peer Prediction to the Test. In *Proceedings of the 15th ACM Conference on Electronic Commerce (EC'14)*, 507–524.

Gneiting, T., and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102:359–378.

Good, I. J. 1952. Rational Decisions. *Journal of the Royal Statistical Society. Series B* 14(1):107–114.

John, L. K.; Loewenstein, G.; and Prelec, D. 2012. Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling. *Psychological Science* 23(5):524–532.

Jurca, R. 2007. *Truthful Reputation Mechanisms for Online Systems*. Ph.D. Dissertation, Ecole Polytechnique Fédérale de Lausanne (EPFL).

Lambert, N. S.; Pennock, D. M.; and Shoham, Y. 2008. Eliciting Properties of Probability Distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08)*, 129–138. ACM.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51(9):1359–1373.

Prelec, D. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306(5695):462–466.

Radanovic, G., and Faltings, B. 2013. A Robust Bayesian Truth Serum for Non-binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*, 833–839.

Rigol, N. 2016. *Essays on the Design and Targeting of Financial Products: Evidence from Field Experiments in India*. Ph.D. Dissertation, Department of Economics, Massachusetts Institute of Technology.

Savage, L. J. 1971. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association* 66:783–801.

Selten, R. 1998. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics* 1:43–61.

Shnayder, V.; Frongillo, R.; and Parkes, D. C. 2016. Measuring performance of peer prediction mechanisms using replicator dynamics. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, 2611–2617.

Waggoner, B., and Chen, Y. 2014. Output Agreement Mechanisms and Common Knowledge. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP'14)*, 220–226.

Witkowski, J., and Parkes, D. C. 2012. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, 1492–1498.

Witkowski, J. 2014. *Robust Peer Prediction Mechanisms*. Ph.D. Dissertation, Department of Computer Science, Albert-Ludwigs-Universität Freiburg.

A Intuition for Extremized Mean

Consider a biased coin that is either positively biased (coming up heads with 80%) or negatively biased (coming up heads with 20%). Both biases are equally likely a priori. Two forecasters are now asked to privately flip the coin once and then provide a probabilistic forecast for a single public coin flip that follows their private flips. Assume both are scored using a proper scoring rule, such as the quadratic rule, and so their reports are properly incentivized. Given this setting, there are only two possible posterior beliefs following one flip, namely $0.68 = 0.2^2 + 0.8^2$ and $0.32 = 0.2 \cdot 0.8 + 0.2 \cdot 0.8$. If one forecaster reports 0.68 and the other reports 0.32, then we have learned nothing and the best forecast for the public flip is still 50%. If both report 0.68, however, the best forecast for the public flip is not 0.68 but roughly $\frac{13}{17} \simeq 0.76$, which is the posterior belief for heads given *two* heads have been observed. (The analogous is true for both reporting 0.32.) In this example, the extremized mean is an unbiased proxy for $\alpha \simeq 1.56$. In general, the optimal choice for α depends on the information overlap between forecasters. For example, full information overlap is given when both forecasters not only observe an i.i.d. flip of the same coin but the same flip. In that case, the optimal α would be 1, i.e., no extremizing. For more detailed explanations, see the work by Baron et al. (2014).